**Правительство Российской Федерации**

**Федеральное государственное автономное образовательное учреждение высшего профессионального образования   
«Национальный исследовательский университет   
«Высшая школа экономики»**

**Факультет бизнес-информатики**

**Программа дисциплины**

**«Анализ неструктурированной информации»**

для направления 080500.68 «Бизнес-информатика» подготовки магистра

для магистерской программы «Бизнес-информатика»

Автор программы:

Бекларян А.Л., abeklaryan@hse.ru

Одобрена на заседании кафедры   
бизнес-аналитики «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2013 г.

Зав. кафедрой Кравченко Т.К. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Рекомендована секцией

УМС «Бизнес-информатика» «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2013 г.

Председатель Таратухина Ю.В. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Москва, 2013

*Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения кафедры – разработчика программы.*

# Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает минимальные требования к знаниям и умениям студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и студентов направления 080500.68 «Бизнес-информатика» подготовки магистра, обучающихся по магистерской программе «Бизнес-информатика».

Программа разработана в соответствии с:

* образовательным стандартом Федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики», уровень подготовки: магистр, утвержденным 26.06.2011;
* рабочим учебным планом университета по направлению 080500.68 «Бизнес-информатика» подготовки магистра для магистерской программы «Бизнес-информатика», специализация «Информационная бизнес-аналитика», утвержденным в 2013 г.

# Цели освоения дисциплины

Целями освоения дисциплины является формирование у студентов комплекса теоретических знаний и методологических основ в области систем анализа неструктурированной информации, а также практических навыков, необходимых для внедрения и практического использования таких систем.

# Компетенции обучающегося формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

* знать характеристики рынка систем анализа неструктурированной информации и перспективы развития сегмента информационно-технологической отрасли, получившей название «Большие данные» (Big Data), основные методы анализа, применяемые в «Больших данных», в том числе А/Б тестирование (A/B Testing), Кластерный анализ (Cluster Analysis), Краудсорсинг (Crowdsourcing), Ансамблевое обучение (Ensemble Learning), Генетические (эволюционные) алгоритмы (Genetic Algorithms), Сетевой анализ (Network Analysis) и др., а также основные классы и принципы построения информационных систем, применяемых для практической реализации этих методов;
* уметь применять системы анализа неструктурированной информации для решения задач сквозного поиска по источникам, выявления закономерностей на основании анализа текстовых данных, извлечения ключевых факторов из неструктурированных текстов;
* иметь навыки (приобрести опыт) сбора данных, обработки текстов на естественном языке, отображения выявленных тенденций, зависимостей, аномалий, основываясь на работе в системе IBM Content Analytics.

В результате освоения дисциплины студент приобретает следующие компетенции:

| Компетенция | Код по ФГОС/ НИУ | Дескрипторы – основные признаки освоения (показатели достижения результата) | Формы и методы обучения, способствующие формированию и развитию компетенции |
| --- | --- | --- | --- |
| Знать теоритические основы моделирования, методологию построения моделей сложных систем, методы сбора, хранения и анализа информации | СК-1  СК-3 | Владеет и применяет | Лекции, практические занятия, выполнение домашних заданий |
| Способность выявлять и прогнозировать основные направления использования современных ИКТ для управления эффективностью бизнеса | ПК-11 | Владеет и применяет | Лекции, практические занятия, выполнение домашних заданий |
| Способность проводить научные исследования и готовить аналитические материалы для оценки мероприятий и выработки стратегических решений в сфере ИКТ | ПК-12 | Владеет и применяет | Лекции, практические занятия, выполнение домашних заданий |
| Способность обосновывать решения на основе результатов моделирования, | СК-2 | Демонстрирует | Лекции, практические занятия, выполнение домашних заданий |
| Владение методами теории систем и системного анализа | ПК-8  ПК-13 | Демонстрирует | Лекции, практические занятия, выполнение домашних заданий |

# Место дисциплины в структуре образовательной программы

Для специализации «Информационная бизнес-аналитика» в рамках магистерской программы «Бизнес-информатика» настоящая дисциплина является дисциплиной по выбору.

Изучение данной дисциплины базируется на следующих дисциплинах:

* «Проектирование информационных систем»;
* «Системный анализ и проектирование»;
* «Управление данными»;
* «Теория вероятностей и математическая статистика».

Для освоения учебной дисциплины, студенты должны знать концептуальные основы архитектуры предприятия, основные классы информационных систем управления бизнесом, лучшие практики и современные стандарты в сфере информационных технологий.

Также студенты должны владеть методами проектирования информационных систем, уметь систематизировать и обобщать информацию, разрабатывать конкретные предложения  по результатам исследований, готовить справочно-аналитические материалы для принятия управленческих решений в сфере информационных технологий.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

* «Системы интеллектуального анализа данных»;
* «Системы бизнес интеллекта»;
* «Системы поддержки принятия решений».

# Тематический план учебной дисциплины

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № | Название раздела | Всего часов | Аудиторные часы | | | Самостоя­тельная работа |
| Лек-ции | Семи-нары | Практи-ческие занятия |
| **1.** | **КОНЦЕПЦИЯ «БОЛЬШИХ ДАННЫХ»** | **32** | **10** |  | **0** | **22** |
| 1.1. | Возникновение феномена «Больших данных» | 8 | 2 |  | 0 | 6 |
| 1.2. | Преобразование текста в аналитический ресурс | 14 | 4 |  | 0 | 10 |
| 1.3. | Технический аспект «Больших данных» | 10 | 4 |  | 0 | 6 |
| **2.** | **НЕСТРУКТУРИРОВАННАЯ ИНФОРМАЦИЯ** | **52** | **10** |  | **4** | **38** |
| 2.1. | Концепция Enterprise Content Management (ЕСМ) | 8 | 2 |  | 0 | 6 |
| 2.2. | Методы анализа неструктурированной информации | 20 | 2 |  | 2 | 16 |
| 2.3. | Обработка естественных языков и анализ настроений | 24 | 6 |  | 2 | 16 |
| **3.** | **АППАРАТНОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ «БОЛЬШИХ ДАННЫХ»** | **52** | **6** |  | **6** | **40** |
| 3.1. | Вычислительная парадигма MapReduce и концепция NoSQL | 28 | 4 |  | 4 | 20 |
| 3.2. | Центры хранения данных и управляющее программное обеспечение для массово-параллельной обработки | 24 | 2 |  | 2 | 20 |
| **4.** | **МАСШТАБИРОВАНИЕ И МНОГОУРОВНЕВОЕ ХРАНЕНИЕ «БОЛЬШИХ ДАННЫХ»** | **38** | **6** |  | **8** | **24** |
| 4.1. | Облачные хранилища и облачные вычисления | 16 | 2 |  | 2 | 12 |
| 4.2. | Быстрые Данные (Fast Data), Большая Аналитика (Big Analytics) и Глубокое Проникновение (Deep Insight) | 22 | 4 |  | 6 | 12 |
| **5.** | **ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ «БОЛЬШИХ ДАННЫХ»** | **42** | **0** |  | **14** | **28** |
| 5.1. | «Большие данные» и анализ интернет контента | 22 | 0 |  | 8 | 14 |
| 5.2. | «Большие данные» и мониторинг общественного мнения | 20 | 0 |  | 6 | 14 |
|  | **Итого часов** | **216** | **32** |  | **32** | **152** |

# Формы контроля знаний студентов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Тип**  **контроля** | **Форма контроля** | **1 год** | | **Параметры** |
| **2** | **3** |
| Текущий  (неделя) | Контрольная  работа |  | 9 | Письменная работа 90 минут, оценка результатов – 1 неделя |
| Реферат |  | 8 | Объем 20-25 стр., оценка результатов – 2 недели |
| Домашнее  задание |  | 9 | Пояснительная записка до 20 стр., оценка результатов – 2 недели |
| Итоговый  (неделя) | Экзамен |  | 10 | 1. Устный экзамен 2. Письменный экзамен на базе материалов IBM. |

## Критерии оценки знаний, навыков

Студент должен продемонстрировать знание разделов дисциплины и способность представить результаты выполнения домашних заданий и качественный реферат на заданную тему в соответствии с требуемыми компетенциями.

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале.

# Содержание дисциплины

###### Раздел 1. КОНЦЕПЦИЯ «БОЛЬШИХ ДАННЫХ»

###### Тема 1.1. Возникновение феномена «Больших данных»

Что такое «Большие данные», и что они нам сулят. Разница между бизнес-аналитикой и «Большими данными». Устаревание информации. Рост объемов данных на фоне вытеснения аналоговых средств хранения. Корректная интерпретация информационных потоков. Обработка информационных потоков. Предпосылки применения контент-анализа в различных исследованиях.

**Основная литература**

Барсегян А.А. Анализ данных и процессов. – БХВ-Петербург, 2009. – с. 512.

**Дополнительная литература**

Тюрин Ю.Н. Анализ данных на компьютере. – М. ИНФРА-М, 2003. – с. 544.

Zikopoulos P.C., Eaton C., De Roos D., Deutsch T., Lapis G. Understanding Big Data. Analytics for Enterprise Class Hadoop and Streaming Data. – McGraw-Hill, 2012.

###### Тема 1.2. Преобразование текста в аналитический ресурс

Необходимость в аналитической работе с большими данными. Явная (выраженная) и скрытая (структурная) информация. Количественная и качественная стратегия анализа текстов. Возможности и ограничения каждого из подходов. Процедура контент-анализа. Определение круга проблем для контент-анализа. Начальный этап исследования: формулирование целей и задач исследования, выбор эмпирического материала, выдвижение рабочих гипотез. Операциональный этап исследования: определение категорий и подкатегорий, выбор единиц анализа, установление правил кодирования. Этап счета. Этап интерпретации результатов. Презентация результатов. Типичные ошибки при проведении контент-анализа.

**Основная литература**

Барсегян А.А. Анализ данных и процессов. – БХВ-Петербург, 2009. – с. 512.

Krippendorff K. Content analysis. – Los Angeles SAGE Publications, 2013. – 440 p.

**Дополнительная литература**

Тюрин Ю.Н. Анализ данных на компьютере. – М. ИНФРА-М, 2003. – с. 544.

Шалак В.И. Современный контент-анализ. – ОМЕГА-Л, 2009. – с. 272.

Weber R. Basic Content Analysis. Newbury Park. Calif., 1990.

###### Тема 1.3. Технический аспект «Больших данных»

Технические признаки, характеризующие «Большие данные». Принцип V3 – Volume (объём данных), Variety (разнообразие данных) и Velocity (скорость генерации и работы с данными). Интеграция, миграция и построение хранилищ данных. Высокопроизводительные вычисления (High Performance Computing, HPC) при выполнении аналитических исследований. Grid computing (распределенные вычисления на нескольких серверах), in-database analytics (частичный перевод нагрузки при аналитических вычислениях в СУБД, а также регламентное применение готовых аналитических моделей к новым данным полностью на стороне СУБД) и in-memory analytics (применение аналитики прямо в оперативной памяти сервера СУБД).

**Основная литература**

Барсегян А.А. Анализ данных и процессов. – БХВ-Петербург, 2009. – с. 512.

**Дополнительная литература**

Методологические и методические проблемы контент-анализа. Вып. 1-2. – М. – Л., 1973.

Big Data: The Next Frontier for Innovation, Competition, and Productivity. – McKinsey Global Institute, May 2011.

Big Data: What It Is and Why You Should Care. White Paper. – IDC, 2011.

Zikopoulos P.C., Eaton C., De Roos D., Deutsch T., Lapis G. Understanding Big Data. Analytics for Enterprise Class Hadoop and Streaming Data. – McGraw-Hill, 2012.

###### Раздел 2. НЕСТРУКТУРИРОВАННАЯ ИНФОРМАЦИЯ

###### Тема 2.1. Концепция Enterprise Content Management (ЕСМ)

Управление корпоративными ресурсами, сопровождение информации, управление документами (создание, контроль, безопасность, возврат и перемещение документа, группировка); управление бумажными документами (в том числе их оцифровка путем сканирования); возможность осуществления групповой работы над документами и объединение их в проекты; организация архивного хранения документов (при этом соблюдаются как законодательные нормы, так и внутренние нормы предприятия). Организация документов, поддерживающих бизнес-процессы предприятия, в том числе, осуществление маршрутизации заданий с дальнейшей проверкой исполнения; управление веб-контентом, который применяется для публикации, а также контентом, который является вспомогательным для работников предприятия.

**Основная литература**

Барсегян А.А. Анализ данных и процессов. – БХВ-Петербург, 2009. – с. 512.

**Дополнительная литература**

Аверьянов Л.Я. Контент-анализ. – М. КноРус, 2009. – с. 451.

Тюрин Ю.Н. Анализ данных на компьютере. – М. ИНФРА-М, 2003. – с. 544.

Шалак В.И. Современный контент-анализ. – ОМЕГА-Л, 2009. – с. 272.

Тезаурус социологии. Книга 2. Методология и методы социологических исследований. Тематический словарь-справочник. Под редакцией: Тощенко Ж.Т. – М.: Юнити-Дана, 2013. – с. 416.

Mancini J. Enterprise Content Management: Critical Technologies for Business Applications // AIIM, 2001.

###### Тема 2.2. Методы анализа неструктурированной информации

Эвристические алгоритмы поиска, эволюционное вычисление, этапы генетического алгоритма: задание целевой функции (приспособленности) для особей популяции, создание начальной популяции, размножение (скрещивание), мутирование, вычисление значения целевой функции для всех особей, формирование нового поколения (селекция).

Задача кластеризации, методы кластеризации, иерархическая кластеризация, алгоритм k-средних, зонтичная кластеризация, методы ненаправляемого обучения (Unsupervised Learning). Постановка задачи классификации, подходы и применения, построение и обучение классификатора, оценка качества классификации, рубрикации тренировочных данных (Training Data Set), методы управляемого (направляемого) обучения (Supervised Learning).

Методы распознавания образов, дискриминантный анализ, нелинейная оптимизация, этапы формирования нейронных сетей: сбор данных для обучения, подготовка и нормализация данных, выбор топологии сети, экспериментальный подбор характеристик сети, экспериментальный подбор параметров обучения, собственно обучение, проверка адекватности обучения, корректировка параметров, окончательное обучение, вербализация сети с целью дальнейшего использования.

**Основная литература**

Барсегян А.А. Анализ данных и процессов. – БХВ-Петербург, 2009. – с. 512.

**Дополнительная литература**

Гладков Л.А., Курейчик В.В., Курейчик В.М. Генетические алгоритмы: Учебное пособие. – 2-е изд. – М: Физматлит, 2006. – с. 320.

Емельянов В.В., Курейчик В.В., Курейчик В.М. Теория и практика эволюционного моделирования. – М: Физматлит, 2003. – с. 432.

Фомичева И.Д. Социология СМИ. Учебное пособие. Рекомендовано УМО по классическому университетскому образованию РФ в качестве учебного пособия для студентов высших учебных заведений. – М. Аспект Пресс, 2012. – с. 360.

###### Тема 2.3. Обработка естественных языков и анализ настроений

Совместное использование компьютерных технологий и лингвистики для создания алгоритмов, позволяющих анализировать естественные (человеческие) языки. Применение методов обработки естественных языков и других аналитических методов для выявления и извлечения из анализируемого текста субъективной информации, характеризующей настроения, мнения, отношение людей к проблеме. Рассмотрение следующих основных задач: синтез речи, распознавание речи, анализ текста, синтез текста, машинный перевод, вопросно-ответные системы, информационный поиск, извлечение информации, анализ тональности текста, анализ высказываний, упрощение текста.

**Основная литература**

Барсегян А.А. Анализ данных и процессов. – БХВ-Петербург, 2009. – с. 512.

Krippendorff K. Content analysis. – Los Angeles SAGE Publications, 2013. – 440 p.

**Дополнительная литература**

Фомичева И.Д. Социология СМИ. Учебное пособие. Рекомендовано УМО по классическому университетскому образованию РФ в качестве учебного пособия для студентов высших учебных заведений. – М. Аспект Пресс, 2012. – с. 360.

Шалак В.И. Современный контент-анализ: Приложения в области: политологии, психологии, социологии, культурологии, экономики, рекламы. – М., 2004.

Методологические и методические проблемы контент-анализа. Вып. 1-2. – М. – Л., 1973.

###### Раздел 3. АППАРАТНОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ «БОЛЬШИХ ДАННЫХ»

###### Тема 3.1. Вычислительная парадигма MapReduce и концепция NoSQL

Вычисления некоторых наборов распределенных задач с использованием большого количества компьютеров, образующих кластер. Шаги Map и Reduce. Предварительная обработка входных данных и свёртка данных. Концепция параллелизма. Шаблоны доступа к данным, хеш-таблица, деревья, таксономия NoSQL, колоночные СУБД, bigtable.

**Основная литература**

Барсегян А.А. Анализ данных и процессов. – БХВ-Петербург, 2009. – с. 512.

**Дополнительная литература**

Фаулер М., Прамодкумар Дж. Садаладж. NoSQL: новая методология разработки нереляционных баз данных. – М.: «Вильямс», 2013. – с. 192.

Dean J., Ghemawat S. MapReduce: Simplified data processing on large clusters. In Proceedings of the Sixth Conference on Operating System Design and Implementation – Berkeley, CA, 2004.

###### Тема 3.2. Центры хранения данных и управляющее программное обеспечение для массово-параллельной обработки

Разработка и выполнение распределённых программ, расширение вычислительных мощностей посредством добавления в кластер дополнительных узлов, технология Hadoop, распределённая файловая система HDFS (Hadoop Distributed File System), интеграция с NoSQL и MapReduce.

**Основная литература**

Барсегян, А.А. Анализ данных и процессов. – БХВ-Петербург, 2009. – с. 512.

**Дополнительная литература**

Лэм Ч. Hadoop в действии. – М.: ДМК Пресс, 2012.

White T. Hadoop: The Definitive Guide. – 2-nd edition. – Sebastopol: O’Reilly Media, 2011. – 600 p.

###### Раздел 4. МАСШТАБИРОВАНИЕ И МНОГОУРОВНЕВОЕ ХРАНЕНИЕ «БОЛЬШИХ ДАННЫХ»

###### Тема 4.1. Облачные хранилища и облачные вычисления

Модели развёртывания: частное облако, публичное облако, гибридное облако, общественное облако. Модели обслуживания: программное обеспечение, платформа, инфраструктура. Экономические аспекты центров обработки данных. Безопасность при хранении и пересылке данных. Проблема «последней мили».

**Основная литература**

Барсегян А.А. Анализ данных и процессов. – БХВ-Петербург, 2009. – с. 512.

**Дополнительная литература**

Гребнев Е. Облачные сервисы. Взгляд из России. Под ред. – М.: CNews, 2011.

Gillam L. Cloud Computing: Principles, Systems and Applications – L.:Springer,2010. – 379 p.

###### Тема 4.2. Быстрые Данные (Fast Data), Большая Аналитика (Big Analytics) и Глубокое Проникновение (Deep Insight)

Обработка Fast Data, подтверждение и корректировка априорных знаний и гипотез, синхронизация скорости работы с ростом объема данных. Получение знаний посредством Big Analytics, преобразования зафиксированной в данных информации в новое знание, принцип «обучения с учителем». Высший уровень работы с данными Deep Insight, обучение без учителя (unsupervised learning), использование современных методов аналитики, а также различные способы визуализации, обнаружение знаний и закономерностей, априорно неизвестных.

**Основная литература**

Барсегян А.А. Анализ данных и процессов. – БХВ-Петербург, 2009. – с. 512.

**Дополнительная литература**

Zikopoulos P.C., Eaton C., De Roos D., Deutsch T., Lapis G. Understanding Big Data. Analytics for Enterprise Class Hadoop and Streaming Data. – McGraw-Hill, 2012.

Big Data: The Next Frontier for Innovation, Competition, and Productivity. – McKinsey Global Institute, May 2011.

Big Data: What It Is and Why You Should Care. White Paper. – IDC, 2011.

Mancini J. Enterprise Content Management: Critical Technologies for Business Applications // AIIM, 2001.

###### Раздел 5. ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ «БОЛЬШИХ ДАННЫХ»

###### Тема 5.1. «Большие данные» и анализ интернет контента

Практическое применение решений IBM Content Analytics. Понятие шаблона, создание правил и категорий. Персональная база данных, фразовый поиск, нечеткий поиск. Возможности уточнения результатов запросов с учетом структуры текста. Анализ совместной встречаемости (collocate analysis) и коэффициент связи категорий (Z-score).

**Основная литература**

Барсегян А.А. Анализ данных и процессов. – БХВ-Петербург, 2009. – с. 512.

**Дополнительная литература**

Аверьянов Л.Я. Контент-анализ. – М. КноРус, 2009. – с. 451.

Тюрин Ю.Н. Анализ данных на компьютере. – М. ИНФРА-М, 2003. – с. 544.

Шалак В.И. Современный контент-анализ: Приложения в области: политологии, психологии, социологии, культурологии, экономики, рекламы. – М., 2004.

**Тема 5.2. «Большие данные» и мониторинг общественного мнения**

Практическое применение решений IBM Content Analytics. Контент-анализ массовой корреспонденции и социологических опросов. Прямые пропорциональные закономерности, аддитивные закономерности, мультипликативные закономерности.

**Основная литература**

Барсегян А.А. Анализ данных и процессов. – БХВ-Петербург, 2009. – с. 512.

**Дополнительная литература**

Аверьянов Л.Я. Контент-анализ. – М. КноРус, 2009. – с. 451.

Тюрин Ю.Н. Анализ данных на компьютере. – М. ИНФРА-М, 2003. – с. 544.

Фомичева И.Д. Социология СМИ. Учебное пособие. Рекомендовано УМО по классическому университетскому образованию РФ в качестве учебного пособия для студентов высших учебных заведений. – М. Аспект Пресс, 2012. – с. 360.

Шалак В.И. Современный контент-анализ: Приложения в области: политологии, психологии, социологии, культурологии, экономики, рекламы. – М., 2004.

# Образовательные технологии

Образовательные технологии, используемые при реализации различных видов учебной работы: доклады, обсуждения, решение задач, рассмотрение кейсов.

# Оценочные средства для текущего контроля студентов

## Содержание заданий текущего контроля

Аудиторная контрольная работа проводится по решению конкретной задачи с применением методов анализа неструктурированной информации.

Выполнение домашних заданий предусматривает построение моделей анализа неструктурированной информации, выявление регулярных выражений, построение аналитических срезов и фильтров, выделение корреляций между срезами, отображение взаимосвязей и визуализацию итогов анализа в системе IBM Content Analytics.

Написание рефератов осуществляется на основе зарубежных (англоязычных) публикаций по вопросам теории и практики систем анализа неструктурированной информации, а также на основе результатов выполнения домашних заданий.

## Тематика рефератов

1. Сравнительный анализ аппаратных решений для «Больших данных».
2. Методы машинного обучения и искусственный интеллект.
3. Персептрон и кибернетическая модель мозга.
4. Искусственные системы распознавания образов.
5. Сравнительный анализ распределенных файловых систем.
6. Сравнительный анализ key-value хранилищ.
7. Сравнительный анализ поколоночных и key-value хранилищ.
8. Машинная обработка естественных языков.
9. Переосмысление глобальных данных: рост объемов и ценности информации (Mapping global data: growth and value creation).
10. Преобразование потенциала больших данных в сфере здравоохранения (The transformative potential of big data in health care domain).
11. Преобразование потенциала больших данных в сфере государственного управления (The transformative potential of big data in public sector administration domain).
12. Преобразование потенциала больших данных в сфере торговли (The transformative potential of big data in retail domain).
13. Преобразование потенциала больших данных в сфере промышленности (The transformative potential of big data in manufacturing domain).
14. Преобразование потенциала больших данных в сфере навигации (The transformative potential of big data in personal location data domain).
15. Влияние «Больших данных» на руководителей организаций (Implications for organization leaders).
16. Влияние «Больших данных» на политику (Implications for policy makers).
17. Построение индексов потенциальной ценности и простоты преобразования информации (Construction of indices on value potential and ease of capture).
18. Методология преобразования информации (Data map methodology).
19. Методология анализа спроса и предложения на специалистов в области бизнес-аналитики (Methodology for analyzing the supply and demand of analytical talent).
20. Hadoop-Hive: SQL-подобное манипулирование данными.
21. Система ввода-вывода в Hadoop (Hadoop I/O).
22. Типы и форматы MapReduce (MapReduce types and formats).
23. Настройка кластера Hadoop (Setting up a Hadoop cluster).
24. Администрирование Hadoop (Administering Hadoop).
25. Pig: платформа создания MapReduce приложений.
26. Нереляционная распределённая база данных HBase.
27. ZooKeeper: централизованная служба поддержки конфигурации.
28. Sqoop: простая миграция из реляционных баз данных.
29. Установка Apache Hadoop (Installing Apache Hadoop).
30. Oozie: планировщик рабочего процесса системы управления рабочими местами.
31. Flume: система для эффективного сбора, объединения и перемещения больших объемов данных, полученных из различных источников, в централизованное хранилище данных.
32. Hue: пользовательский интерфейс Hadoop.
33. Avro: удаленный вызов процедур и сериализация.
34. Таксономия, исследование и другие вопросы экосистемы облачных вычислений (A taxonomy, survey, and issues of cloud computing ecosystems).
35. Изучение облачных вычислений с точки зрения сетевого анализа и совместной работы (Examining cloud computing from the perspective of grid and computer-supported sooperative work).
36. Анализ облачных стандартов (Overview of cloud standards).
37. P2P фреймворк для поддержки приложений MapReduce в динамических облаках (A peer-to-peer framework for supporting MapReduce applications in dynamic cloud environments).
38. Улучшенная поддержка сети для масштабируемых облачных вычислений (Enhanced network support for scalable computing clouds).
39. YML-PC: эталонная архитектура на основе workflow для создания частных наукоемких облаков (YML-PC: a reference architecture based on workflow for building scientific private clouds).
40. Управление ресурсами для гибридных сетей и облачных вычислений (Resource management for hybrid grid and cloud computing).
41. P2P расшаривание облака: эффективный поиск и распределение нагрузки (Peer-to-Peer cloud provisioning: service discovery and load-balancing).
42. Облачные вычисления – конфиденциальность данных и проблемы совместимости (Cloud computing – data confidentiality and interoperability challenges).
43. Технологии для исполнения и распространения правил и политик в облачной архитектуре (Technologies for enforcement and distribution of policy in cloud architectures).

## Вопросы для оценки качества освоения дисциплины

###### Вопросы к Разделу 1. КОНЦЕПЦИЯ «БОЛЬШИХ ДАННЫХ»

1. В чем принципиальное отличие концепции Big Data от традиционного подхода BI?
2. Понятие явной (выраженной) и скрытой (структурной) информации.
3. Определение контент-анализа.
4. Каковы основные понятия контент-анализа?
5. Какие существуют виды контент-анализа?
6. Какие существуют этапы контент-анализа?
7. Каковы основные признаки, характеризующие «Большие данные»?

###### Вопросы к Разделу 2. НЕСТРУКТУРИРОВАННАЯ ИНФОРМАЦИЯ

1. Концепция Enterprise Content Management (ЕСМ).
2. Сущность и задачи кластеризации.
3. Основные понятия, принципы и предпосылки генетических алгоритмов.
4. Достоинства и недостатки генетических алгоритмов.
5. Классификация нейронных сетей и принципы построения.
6. Искусственная нейронная сеть прямого прохода.
7. Использование генетических алгоритмов для обучения искусственных нейронных сетей
8. Кластеризация как инструмент предварительной обработки данных для искусственной нейронной сети
9. Какова цель синтаксического анализа?
10. Общая схема алгоритма синтаксического анализа «сверху-вниз» и «снизу-вверх».

###### Вопросы к Разделу 3. АППАРАТНОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ «БОЛЬШИХ ДАННЫХ»

1. Схема работы фаз map(ƒ, c) и reduce(ƒ, c).
2. Преимущества, ограничения и недостатки парадигмы MapReduce.
3. Какие бывают модели данных и запросов в NoSQL?
4. Какие бывают системы хранения данных в NoSQL?
5. Основные принципы работы фреймворка Hadoop.
6. Репликация данных в распределенной файловой системе HDFS.

###### Вопросы к Разделу 4. МАСШТАБИРОВАНИЕ И МНОГОУРОВНЕВОЕ ХРАНЕНИЕ «БОЛЬШИХ ДАННЫХ»

1. Модели развертывания облачных хранилищ.
2. Модели обслуживания облачных хранилищ.
3. Постановка и описание проблемы «последней мили».
4. Безопасность, производительность и надежность при работе с облачными данными.
5. Экономическая составляющая облачных подходов.
6. Способы машинного обучения.
7. Основные фазы обработки «больших данных».

###### Вопросы к Разделу 5. ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ «БОЛЬШИХ ДАННЫХ»

1. Чем отличаются текстовая и персональная базы данных?
2. Метод анализа комбинации слов (collocate analysis).
3. Понятие "сила связи".
4. Статистическая мера совместной встречаемости слов и категорий (Z-score).
5. Реализация закономерностей в системе IBM Content Analytics.

# Порядок формирования итоговой оценки по дисциплине

Формирование оценок по учебной дисциплине производится в соответствии с Положением об организации контроля знаний, утвержденного Ученым советом НИУ ВШЭ от 24.06.2011, протокол №26.

В соответствии с Рабочим учебным планом, формами текущего контроля являются контрольная работа, реферат и домашнее задание. Каждая из форм текущего контроля оценивается по 10-балльной шкале. Общая оценка за текущий контроль (по 10-балльной шкале) рассчитывается по формуле:

*Отекущий* = *0,2 · Ок/р + 0,5 · Ореф + 0,3 · Одз* ,

где *Ок/р* – оценка за контрольную работу;

*Ореф* – оценка за реферат;

*Одз* – оценка за домашнее задание.

При определении накопленной оценки (по 10-балльной шкале) аудиторная работа и самостоятельная внеаудиторная работа не оцениваются. Поэтому накопленная оценка совпадает с оценкой за текущий контроль и рассчитывается по формуле:

*Онакопленная* = *1,0 · Отекущий + 0,0 · Оауд + 0,0 · Осам.работа* ,

где *Отекущий* – оценка за текущий контроль;

*Оауд* – оценка за аудиторную работу;

*Осам.работа* – оценка за самостоятельную работу.

Результирующая оценка (выставляется в диплом) формируется на основе итоговой оценки за экзамен (по 10-балльной шкале) и накопленной оценки. Результирующая оценка рассчитывается по формуле:

*Орезульт* = *0,3 · Оэкз + 0,7 · Онакопленная* ,

где *Оэкз* – оценка за итоговый контроль (экзамен);

*Онакопленная* – накопленная оценка.

При формировании оценок на основе весовых коэффициентов применяется округление до целого числа в большую сторону.

# Учебно-методическое и информационное обеспечение дисциплины

## Базовый учебник

Барсегян А.А. Анализ данных и процессов. – БХВ-Петербург, 2009. – с. 512.

## Основная литература

Аверьянов Л.Я. Контент-анализ. – М. КноРус, 2009. – с. 451.

Тюрин Ю.Н. Анализ данных на компьютере. – М. ИНФРА-М, 2003. – с. 544.

Фомичева И.Д. Социология СМИ. Учебное пособие. Рекомендовано УМО по классическому университетскому образованию РФ в качестве учебного пособия для студентов высших учебных заведений. – М. Аспект Пресс, 2012. – с. 360.

Шалак В.И. Современный контент-анализ. – ОМЕГА-Л, 2009. – с. 272.

Krippendorff K. Content analysis. – Los Angeles SAGE Publications, 2013. – 456 p.

Zikopoulos P.C., Eaton C., De Roos D., Deutsch T., Lapis G. Understanding Big Data. Analytics for Enterprise Class Hadoop and Streaming Data. – McGraw-Hill, 2012.

## Дополнительная литература

Гладков Л.А., Курейчик В.В., Курейчик В.М. Генетические алгоритмы: Учебное пособие. – 2-е изд. – М: Физматлит, 2006. – с. 320.

Гребнев Е. Облачные сервисы. Взгляд из России. – М.: CNews, 2011.

Емельянов В.В., Курейчик В.В., Курейчик В.М. Теория и практика эволюционного моделирования. – М: Физматлит, 2003. – с. 432.

Лэм Ч. Hadoop в действии. – М.: ДМК Пресс, 2012.

Методологические и методические проблемы контент-анализа. Вып. 1-2. – М. – Л., 1973.

Фаулер М., Прамодкумар Дж. Садаладж. NoSQL: новая методология разработки нереляционных баз данных. – М.: «Вильямс», 2013. – с. 192.

Шалак В.И. Современный контент-анализ: Приложения в области: политологии, психологии, социологии, культурологии, экономики, рекламы. – М., 2004.

Dean J., Ghemawat S. MapReduce: Simplified data processing on large clusters. In Proceedings of the Sixth Conference on Operating System Design and Implementation – Berkeley, CA, 2004.

Gillam L. Cloud Computing: Principles, Systems and Applications – L.:Springer,2010. – 379 p.

Weber R. Basic Content Analysis. Newbury Park. Calif., 1990.

White T. Hadoop: The Definitive Guide. – 2-nd edition. – Sebastopol: O’Reilly Media, 2011. – 600 p.

## Справочники, словари, энциклопедии

Тезаурус социологии. Книга 2. Методология и методы социологических исследований. Тематический словарь-справочник. Под редакцией: Тощенко Ж.Т. – М.: Юнити-Дана, 2013. – с. 416.

Big Data: The Next Frontier for Innovation, Competition, and Productivity. – McKinsey Global Institute, May 2011.

Big Data: What It Is and Why You Should Care. White Paper. – IDC, 2011.

## Программные средства

Для подготовки практических задач, докладов и выступлений студентами используется современная учебно-лабораторная база, в том числе:

* стандартные пакеты прикладных программ офисного назначения, в том числе:
  + информационные системы подготовки текстов (Microsoft Word);
  + системы электронных таблиц (Microsoft Excel);
  + системы подготовки презентаций (Microsoft PowerPoint);
* профессиональные информационные системы, в том числе:
  + IBM Cognos BI;
  + IBM Content Analytics;
  + Microsoft SQL Server.

## Дистанционная поддержка дисциплины

Не предусмотрена.

# Материально-техническое обеспечение дисциплины

Используются персональный компьютер (ноутбук) и проектор для проведения лекций и практических занятий, техническое оснащение компьютерных классов.

Авторы программы: **/**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **/** А.Л.Бекларян